第 ** 卷 第 * 期　　中文信息学报　　Vol. **, No. *
201* 年 * 月　　JOURNAL OF CHINESE INFORMATION PROCESSING　　***. , 201*

文章编号：1003-0077（2018）00-0000-00

面向神经机器翻译的集成学习方法分析

李北　王强　姜雨帆　张哲旸　刘继强　张俐　肖桐
（东北大学 自然语言处理实验室，辽宁 沈阳 110819）

摘要：集成学习是一种联合多个学习器进行协同决策的机器学习方法，应用在机器翻译任务的推断过程中可以有效整合多个模型预测的概率分布，达到提升翻译系统的准确性的目的。虽然该方法有效性已在机器翻译评测得到了广泛验证，但关于子模型的选择与融合的策略仍鲜有研究。该文主要针对机器翻译任务中的参数平均与模型融合两种集成学习方法进行了大量的实验，分别从模型与数据层面、多样性与模型数量层面对集成学习的策略进行了深入探索。最终实验结果在 WMT 中英新闻任务上，相比 Transformer 单模型有 3.2 个 BLEU 值的提升。

关键词：神经机器翻译；集成学习；参数平均；模型融合
中图分类号：TP391　　　**文献标识码**：A

On Ensemble Learning of Neural Machine Translation

LI Bei, WANG Qiang, JIANG Yufan, ZHANG Zheyang, LIU Jiqiang, ZHANG Li and XIAO Tong
(NLP Laboratory, Northeastern University, Shenyang, Liaoning 110819, China)

Abstract：Ensemble is a machine learning method that combines multiple learners to make collaborative decisions, which can effectively integrate the probability distributions predicted by multiple models in machine translation tasks, thus improving translation accuracy. Although it has been extensively proved valid in machine translation evaluation, the sub-model selection and integration strategies are still rarely researched. This paper mainly analyzes the two kinds of ensemble learning methods: parameter averaging and model fusion in machine translation tasks. We further explore the impact of diversity and model quantity on system performance from the perspectives of data and model. Experimental results show that the best result yields improvements of 3.2 BLEU points over the strong Transformer baseline on WMT Chinese-English MT tasks.

Key words: Ensemble; Parameter averaging; Model fusion; Diversity

0　引言

集成学习（Ensemble learning）是一种联合多个学习器进行协同决策的机器学习方法[1]。集成学习方法通过整合多个学习器的决策结果可以有效地减小预测结果的方差与偏置[2-4]，显著地提升了模型的泛化能力，达到了比单学习器更好的效果。因此集成学习方法受到了研究人员的广泛认可，被应用于各种实际任务，如规则提取[5]、手写识别[6]等分类回归任务中。

近年来集成学习方法在机器翻译领域也取得了杰出的效果[7]。常见的手段包括平均单模型在训练过程中不同时刻保存的模型参数；在预测过程中整合不同模型的预测结果等。通过在大规模数据的机器翻译评测比赛中进行实验，使用集成学习的手段能大幅度提升翻译的性能，在 CWMT、WMT[8-11]等评测比赛得到了广泛的验证。影响集

收稿收稿日期：2018-06-10；**定稿日期**：2018-07-15
基金项目：国家自然科学基金重点项目(61432013；61732005)；中央高校基本科研业务费；辽宁省高等学校创新人才支持计划
通讯作者：肖桐（xiaotong@mail.neu.edu.cn）

成学习效果的主要因素是模型之间的多样性，因此如何增大模型之间的多样性是提升翻译性能的关键。但大部分研究人员只是在比赛中通过使用集成学习的方法达到提升翻译性能的目的，很少去系统地总结如何才能增大模型间的多样性，缺乏经验性的对比分析与完备性的结论。譬如如何选取具有差异性的子模型，集成多少个模型效果最优等问题并没有得到回答。

针对以上问题，本文基于 Transformer[12]系统分别从模型的参数与解码预测过程两个角度详细总结更高效的 Ensemble 方法。从模型的参数角度，我们将不同时刻保存的模型进行参数平均从而得到更具鲁棒性的单模型；从解码预测过程我们从模型多样性层面、数据多样性层面阐述如何才能增加模型间的多样性来提升翻译的性能。在模型多样性层面，本文主要对比简单的随机初始化种子、复杂种子以及混合不同模型结构的 Ensemble 系统之间的优劣。在数据层面，本文提出分别使用 fine-tuning[13]与 bagging[14]的方式生成子模型，在增加数据多样性的同时提升模型间的差异性。基于以上两方面的讨论，我们又探索了是否使用更多的模型参与融合能在翻译性能上带来正向作用。

本文通过在 WMT17 中英任务的大规模数据集进行实验，我们发现增大模型差异性与数据差异性均能提高模型的翻译性能。此外，随着融合模型的数量增加，翻译质量也会显著地提升。本文最好结果在 Transformer 单模型基础上取得 3.19 个 BLEU 值的提升，并针对模型间多样性与解码复杂度进行了分析。

1 背景

神经机器翻译系统根据给定的源语句子，$x^1,\ldots,x^n$，和目标语句子$y^1,\ldots,y^m$构建模型，建立从源语到目标语的映射函数。这个映射函数通常表示成条件概率$p(y|x)$的形式，翻译的越准确，概率值越高，表示成对数形式：

$$logp(y|x)=\sum_{t=1}^{m}logp(y_t|y_{<t},s) \qquad (1)$$

在 Transformer 提出之前，大多数神经机器翻译系统都是基于注意力机制[15]的循环神经网络（RNN），这些系统虽然取得了不错的效果，但是复杂、循环的计算导致整个训练过程十分漫长，这也是研究者们一直想要改进的地方。随着 Transformer 的提出，它放弃了传统的循环神经网络结构，采用完全基于自注意机制的结构，提高了模型的并行程度，大幅提升了训练速度。同时 Transformer 具有很强的模型表示能力，在翻译的准确性上也有一定的提升。下面来介绍 Transformer 的结构：

Transformer 基于 Encoder-Decoder 结构[16]，采用了一种全新的注意力机制包含 Encoder 端、Decoder 端的自注意机制和 Encoder-Decoder 的联合注意力机制。在机器翻译中编码器负责将输入的源语转换成带有语义信息向量，解码器负责根据语义信息产生目标语句。由于模型没有任何循环或者卷积，为了使用序列的顺序信息，需要将相对以及绝对位置信息注入到模型中去。

编码器由 N 个相同的层堆叠而成，每层都有两个子层，第一个子层是多头自注意力机制，第二个子层是一个简单的（位置敏感的）前馈神经网络。解码器同编码器类似，在编码器的基础上多了一个子层是负责处理编码器输出的多头注意力机制。为了更易于梯度的传递与加速模型收敛，在编码器和解码器中每一个子层后都会有残差连接和层正则化操作[17]。Transformer 使用的多头注意力机制(Mulit-Head Attention)的基本单元是缩放的点积注意力模型(Scaled Dot-Product Attention)，模型结构如图 1 所示，这种注意力的输入是d_k维的 query，d_v维的 key 和 value。具体的计算公式如下：

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (2)$$

其中，Q 中的每一行对应一个 query；使用每个 query 与K^T相乘并进行归一化之后，即可得到一个 n 维的权值向量，然后，再将这个权值向量与 V 相乘，即可得到一个d_v维的加权求和的结果。多头注意力机制就是将 query、key 和 value 映射为 *h* 组维度为d_q、d_k、d_v的向量，分别进行按比例的点积注意力，最后将得到 h 个向量连接起来作为输出。表示如下：

$$head_i=Attention(QW_j^Q,KW_j^K,VW_j^V) \qquad (3)$$

$$MultiHead(Q,K,V)=Concat_j(head_j)W^o \qquad (4)$$

其中，W_j^Q、W_j^K、W_j^V、W^o为参数矩阵。

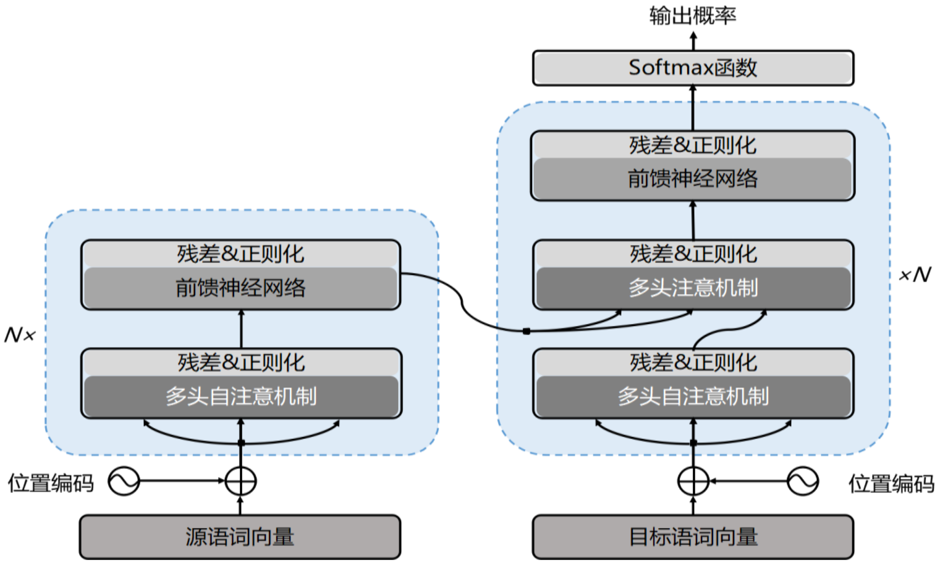

图 1 Transformer 模型架构图

2 Ensemble 方法

本文介绍 Ensemble 在神经机器翻译中的两种应用手段，模型参数平均与预测结果融合。本章节首先介绍如何通过不同时刻保存的模型得到更具鲁棒性的单模型，之后讲述利用参数平均后的各个模型进行预测结果融合过程。在讨论预测结果融合方面分别从模型差异性角度与数据差异性角度来对比分析不同策略带来的增益效果。在此基础上进一步分析通过更多的模型来进行预测结果融合能否带来更大的增益。

2.1 模型参数平均

参数平均是指将单一模型的最近保存的 N 个模型的参数矩阵进行平均。正如图 2 给出 4 个检查点的部分参数矩阵，通过将对应位置数值进行平均得到新的参数矩阵。如模型 1 中的A_{11}=0.26，模型 2 中的B_{11}=0.22，不同检查点（checkpoint）矩阵的对应数值会有差异性，最终经过参数平均后的模型的E_{11}=0.25。Sennrich[8]等人在 WMT16 比赛首次使用模型保存的最新 4 个模型进行模型参数平均的方法，得到了显著地提升。这是由于模型在训练过程中要更新一定的轮数才能达到收敛，并且模型的损失值在收敛后仍处于小范围的上下波动。为了得到更具有鲁棒性的模型，Vaswani[12]等人建议每隔 10 分钟保存一次模型，并平均最新保存的 20 个检查点模型用来作为最终的模型，没给出经验性的结论。本文将要探索如何设置合理的模型保存间隔与参数平均的模型数量获得更强的翻译性能。

2.2 预测结果融合

预测结果融合是一种在解码过程中实施的手段，通过整合不同模型得到的概率分布从而获得新的解，进而预测下一个目标端词语。常见的融合手段有算术平均、几何平均、加权平均以及投票等。机器翻译任务是一种序列生成问题，解码的每一个时序都会依赖于前一时序预测的结果，模型会根据当前的语义信息计算出一个维度大小是词表大小的概率分布向量，经过 Softmax 操作得到归一化的向量表示。向量中每一个元素指代预测下一个词的概率。如图 3 在输入源语“今天 天气 晴朗 。”前提下，解码第一步将 4 个不同模型预测的概率进行算数平均从而得到新的概率分布，其中“today”是当前预测概率最大的词。通过这种方式 Ensemble 模型可以综合不同模型的决策结果来得到更正确的解。在本节主要分别从

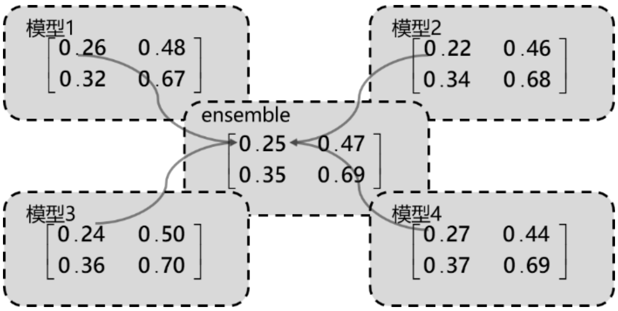

图 2 参数平均结构图

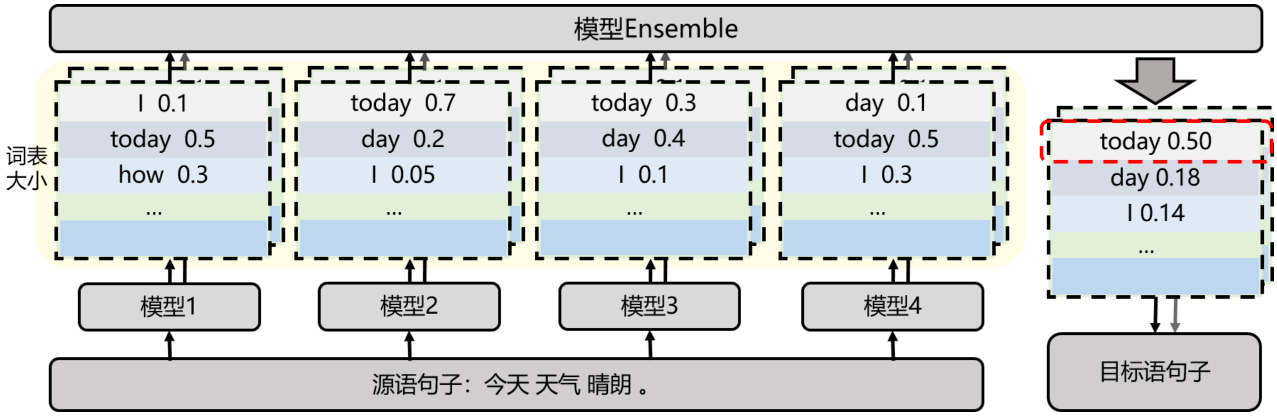

图 3 模型融合示意图

模型多样性与数据多样性的角度来对比分析不同融合的手段带来性能的增益。

2.2.1 模型多样性

模型间的差异性很大程度上决定了模型融合后的效果。对于模型层面，我们将从以下两种策略来构造子模型：

（1）使用相同的模型结构，不同的参数初始化分布或者不同的随机初始化种子。由于神经网络的拟合训练容易陷入局部最优解，最终单模型的翻译结果可能不是全局最优解。通过 Ensemble 的方式，融合多个不同随机种子即多个局部最优解，进而避免这类问题，获得更好的结果。

（2）使用不同的模型结构，不同的随机初始化种子。由于仅仅改变模型的初始化参数，融合相似的网络模型结构差异性过小，为了增大模型间的差异性，本文训练 N 个不同网络结构且不同的随机初始化种子的模型。

2.2.2 数据多样性

在数据层面，为了增加多样性，我们通过 fine-tuning 与 bagging 两种策略来构造子模型：

（1）通过 fine-tuning 的方式在训练好的模型基础上，用分割成不同份数的训练语料继续训练 5 轮，保证词表不变。通过这种方式让微调后的子模型对不同的数据敏感，从而增加了多样性。由于微调的代价很小，不需要完全重新训练模型，节约了很多时间上与设备上的开销。

（2）通过 bagging 的方式对训练样本进行重采样，如图 4 所示分别使用重采样后的子样本训练相应的模型，在预测过程中对不同模型的输出结果上通过取平均的方式得到 ensemble 的模型输出。Bagging 是最早出现在集成学习任务中的有效手段，在早期的分类任务中，通过 bagging 方式构建多个决策树共同对测试数据进行决策，最后采用投票或者取平均的方式得到集成结果。在本任务中通过对训练数据做了 N 次 Bootstrap 采样得到的 N 个训练集近似服从同一分布，同时有效地降低了方差。由于不同子模型之间数据量相同但内容略有差别，从而增加了训练数据的多样性。

2.2.3 更多模型

通过融合更多的模型联合决策来提高翻译性能。大部分研究人员在进行集成方面的研究都只使用了 4 个或 5 个模型进行融合，并没有进一步探究融合更多模型是否有效。本节主要探索参与预测结果平均的模型数量对性能的影响，验证参与融合的模型数量与 Ensemble 模型的翻译性能之间的正相关性。

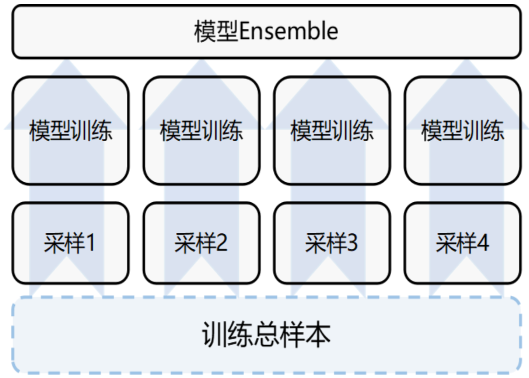

图 4 bagging 示意图

3 实验

3.1 数据筛选

本文实验基于 2018 年 WMT（Workshop on Machine Translation）中英比赛发放的训练数据，其中中英双语平英语料 16M，英文单语数据 24M。

除此之外同时使用了 2018 CWMT（China Workshop on Machine Translation）的双语平行语料，共 9M 句。我们通过以下几个策略对训练语料进行了过滤选择：

-对双语平行语料进行乱码过滤，剔除混有乱码的语料如控制字符、UTF-8 转码生成的单字节乱码等。

-对双语平行语料进行 Niutrans[18]分词处理，保留目标语英文单词的大小写敏感，对中文端的标点符号进行全角转半角操作。

-对双语语料进行长度比过滤，过滤源语端与目标语端超过 100 个词的句子，同时保证源语与目标语长度比在 0.4~3.0 范围内。

-应用 fast-align 脚本对大规模双语语料做词对齐学习，在此基础上生成中英互译词典。清洗源语与目标语端词典覆盖率小于 0.3 的双语语料。

-使用过滤后的双语语料的英文单语训练语言模型[19]，根据语言模型筛选提供的英文单语语料。通过 back-translation 方式生成伪数据[20]，用作数据增强。

-混合筛选的双语平行语料与生成的伪数据，做去重操作。

经过以上几个过滤步骤我们保留了将近 12M 双语平行语料与 4M 的伪数据作为模型的训练数据。我们通过 BLEU 来衡量翻译质量，采用 Moses 提供的 multibleu.perl[21]脚本来计算 BLEU。我们对训练数据进行 BPE 处理[22]，有效的减少 UNK 的出现频次，BPE 的词表大小为 32k，我们的源语端训练词表大小为 48k，目标端训练词表大小为 33k。

3.2 实验设置

本文实验基于 Transformer 模型框架，在 Tensorflow 版的 tensor2tensor 开源工具基础上进行改进。实验的基线设置采用了论文中的 transformer_base 参数设置，编码端与解码端分别是 6 层，隐藏层维度为 512，采用 8 头设置，前馈神经网络的 filter_size 大小为 2048 ，batch-size 大小为 4096 ，采用 8gpu 训练，初始学习率为 0.001，warmup_steps 大小为 4000，采用 Adam 优化器[23]进行参数优化，训练的轮数为 10 轮，共 140K 更新次数。在解码阶段我们采用 beam-search[1]策略来进行解码端的预测，beam-size 大小为 12，同时解码时的长度惩罚 alpha[12]大小为 1.2。基于该参数的单模型在测试集上的 BLEU 值为 25.4，是一个很强的基线设置。

3.3 模型参数平均结果分析

本实验基于 baseline 模型的参数设置，每隔 5min 保存一次模型。基于模型的最新的 20 个 checkpoint，通过对比单模型与 averaging5、averaging10、averaging15、averaging20 之间的性能差异，如图 5 所示：我们发现基于保存的 20 个 checkpoint 点进行参数平均均有一定的正向作用，其中取最新的 15 个模型效果最佳，比最后保存的单一模型高 0.82 BLEU 值。基于本次实验结果，之后的实验均采用参数平均 15 个 checkpoint 后的模型。

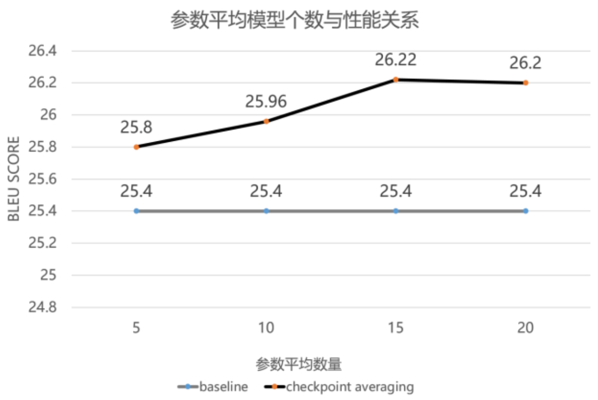

图 5 不同 checkpoint 进行参数平均结果

3.4 预测结果融合结果比较分析

在本章节我们将分别在模型多样性与数据多样性层面讨论预测结果融合的效果。在模型多样性方面，我们主要对比分析在集成的模型数量固定的条件下，相同结构与不同结构两种融合方式的优劣；在数据多样性方面，我们主要探索 fine-tuning 与 bagging 两种手段是否适用于神经机器翻译任务。之后我们会基于以上两类实验结果进一步分析融合数量更多、模型结构差异更大的模型对翻译性能的影响。

3.4.1 模型多样性实验结果

相同模型结构，不同初始化种子 本实验分别采用三种不同的模型设置进行随机种子的实验，每个模型在 WMT17 中英测试集上的 BLEU 表现如图 6 所示，我们发现在 base 参数设置的基础上增大 filter_size 至 4096 与加入 dropout 都会明显提高单模型的 BLEU 分数。通过对相同

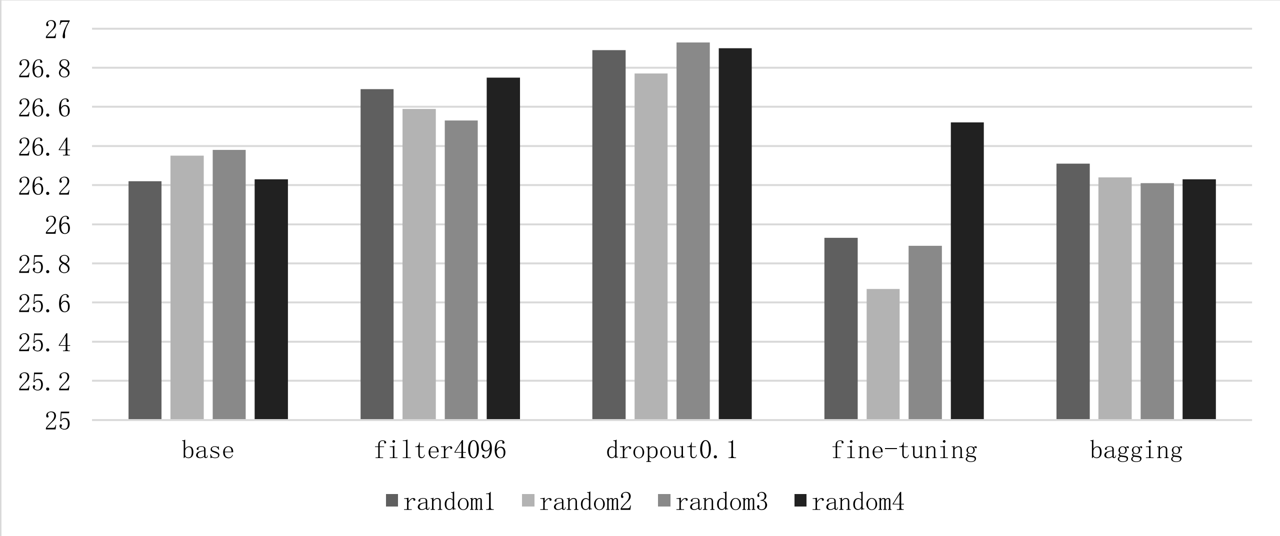

图 6 组间是不同的模型结构，组内是不同的随机初始化种子

模型结构的 4 个不同随机初始化种子进行融合，结果如表 2 中的前三行所示，我们发现三组实验对比于基线在性能方面都有至少+1.5BLEU 的提升。其中增大 filter_size 的 Combo-2 与增加 dropout 的 Combo-3 结果要比 Combo-1 有更好的翻译效果，表明参与集成的基模型性能越强，模型融合后的性能也会上升。另一方面根据 Combo-3 实验结果，增加 dropout 的单模型性能照比 filter4096 有近 0.3BLEU 提升，但模型融合的结果反而要略低于 Combo-2。这里我们猜想是由于 dropout 本身蕴含着集成学习的思想，因此融合 4 个 dropout 模型的性能上升的幅度会有所下降。

不同模型结构，不同初始化种子 本实验主要为了对比更多样化的模型结构能否带来更大的性能提升。为了增加模型间的差异性，我们在 transformer_base 参数设置的基础上

a）加入注意力模型与前向网络的 dropout

b）调整 dropout 大小至 0.2

c）将 filter_size 由 2048 增大至 4096

d）使用 transformer_big 参数设置，由 8 头增加至 16 头、隐层大小由 512 增加至 1024、filter_size 由 2048 增加至 4096、residual_dropout 由 0.1 增大至 0.3

e）引入相对位置表示

根据表 1 的实验结果我们发现增加注意力与激活函数的 dropout 与增大前馈网络的 filter_size 以及使用相对位置均会提高模型的翻译性能。如表 2 中 Combo-4 与 Combo-5 的 BLEU 值比基于 base 模型结构的 Combo-1 的 BLEU 要高 0.4，此外，考虑到参与模型融合的各个单模型之间的性能差异，观察实验 Combo-5 比 Combo-2 仍高近 0.2 BLEU，从而验证了模型间的差异性越大，模型融合的结果会越强。

3.4.2 数据多样性实验结果

使用 fine-tuning 构建子模型 本实验主要在保持源语端与目标语端词表前后一致的前提下，从训练语料中采样出四份不同的训练数据，每份约 4M 句子，分别在 baseline 模型的基础上，对其进行继续训练。由于 baseline 模型已经在原有的 16M 训练集上达到收敛效果，我们从最新的模型保存点继续训练 5 轮左右，让 baseline 模型对不同的训练数据保持敏感，从而增加了 Ensemble 模型之间的差异性。如图 6 所示，在原有 baseline 模型基础上进行微调，大部分模型会在测试集上略有下降，但 Finetune4 模型比 baseline 高近 0.4BLEU。造成这种现象的原因是由于 fine-tuning 过程中数据是随机从原始训练数据中采样的，与测试集句子相近的训练集会直接导致模型的性能上升。根据这一实验现象我们使用测试集的源语端数据训练语言模型，在训练数据中挑选与测试集相近的训练集，每个模型的 BLEU 都在 0.1~0.2 上升区间内。在 fine-tuning 后的模型基础上进行模型融合实验发现 BLEU 值上升了 0.2。这一实验现象验证了随着 ensemble 子模型性能的提升，模型融合后的性能也会有一定的提升。考虑到训练周期与硬件代价问题，通过 fine-tuning 方式构造子模型是一个高效的手段。

使用 bagging 构建子模型 本实验主要在 16M 训练数据的基础上，重采样（无放回）出 4 份总数据量百分之 80 的子样本。基于每个样本使用 baseline 参数设置训练出 4 个模型，每个模型在 WMT17 测试集上的表现如图 6 所示，从实验

结果观测到各个子模型的翻译性能与基线系统相近，没有明显的下降现象。通过模型融合之后的结果见表 2 中的 Combo-7,照比基线提升 1.2 BELU。对比 Combo-1 实验结果，使用 bagging 手段进行模型融合的结果提升幅度要略小一些，可是翻译模型对小数据量的变化不敏感所导致。

3.4.3 更多样的模型

结合模型与数据多样性的模型融合结论，在本章节验证使用更多、差异性更大的模型，是否会在翻译性能上带来正向效果。我们分别融合了 4 个、8 个、12 个、16 个模型进行相应的对比实验。图 7 实验结果表明，参与模型融合的子模型数量与其翻译性能是正相关的。在增大参与预测结果融合的子模型数量的同时，翻译性能整体也会保持上升的趋势。值得注意的是当模型数量上升到一定临界值时，上升的幅度会变小，甚至会略微下降的现象。导致这种现象的原因是伴随着模型融合的数量上升，子模型间的差异性越来越小，因此在增加模型数量的同时也要兼顾模型之间的多样性，我们将在 3.5 节介绍多样性的评价指标。为了解决这个问题，我们通过调整网络的超参数，同时调整 fine-tuning 的手段来进一步增加其模型间的多样性，最终融合 12 个子模型取得了最好的翻译结果，如表 2 中 Combo-11 所示，在测试集上的 BLEU 值为 28.59，对比 Combo-1 高出 0.9 个 BLEU 值。同时，观察集成 8 个模型的 3 组不同实验我们发现，模型之间的差异性越大，融合后提升的效果越显著。

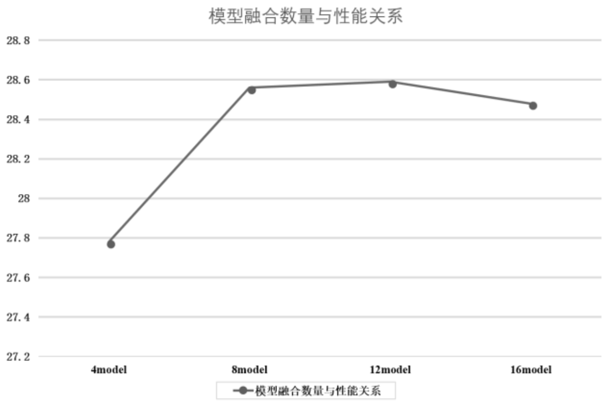

图 7 模型融合数量与性能之间的关系

表 1WMT17-test 不同参数的单模型结果

系统名称	系统描述	Test
baseline	Single model	25.4
Base1	Baseline	26.22
Base2	Random seed	26.35
Base3	Random seed	26.38
Base4	Random seed	26.23
Dropout1	Baseline + attention_dropout = 0.1 + relu_dropout = 0.1	26.89
Dropout2	Baseline + attention_dropout = 0.1 + relu_dropout = 0.1	26.77
Dropout3	Baseline + attention_dropout = 0.1 + relu_dropout = 0.1	26.93
Dropout4	Baseline + attention_dropout = 0.1 + relu_dropout = 0.1	26.90
Base4K1	Baseline + filter_size=4096 + epoch10	26.69
Base4K2	Baseline + filter_size=4096 + epoch10	26.59
Base4K3	Baseline + filter_size=4096 + epoch10	26.53
Base4K4	Baseline + filter_size=4096 + epoch10	26.75
Finetune1	Baseline + 4M + fine-tuning 5epoch	25.93
Finetune2	Baseline + 4M + fine-tuning 5epoch	25.67
Finetune3	Baseline + 4M + fine-tuning 5epoch	25.89
Finetune4	Baseline + 4M + fine-tuning 5epoch	26.52
Bagging1	Bagging13M + transformer_base	26.31
Bagging2	Bagging13M + transformer_base	26.24
Bagging3	Bagging13M + transformer_base	26.21
Bagging4	Bagging13M + transformer_base	26.23
Big	Transformer_big	26.43
Base4K5	Base4K + attention_dropout = 0.1 + relu_dropout = 0.1	27.00
Base4K6	Base4K + attention_dropout = 0.2 + relu_dropout = 0.2	26.83
Finetune5	Baseline + 5M LM + fine-tuning 10epoch	26.67
Base5	Baseline+epoch20	26.58
Rpr4K	Relaive Positon Representation + filter_size=4096	26.98
Base4K7	Base4K + attention_dropout = 0.1 + relu_dropout = 0.1	26.95
Base4K8	Base4K + attention_dropout = 0.1 + relu_dropout = 0.1	26.89

表 2 不同策略模型融合实验在 WMT17-test 测试集结果

系统名称	系统设置	模型数量	Test
Baseline	Transformer_base，Paramter_avg15	1	26.22
Combo-1	Base1,Base2,Base3,Base4	4	27.70
Combo-2	Base4K1,Base4K2,Base4K3,Base4K4	4	27.99
Combo-3	Dropout1, Dropout2, Dropout3, Dropout4	4	27.90
Combo-4	Base5 ,Base4K1,Base4K5, Base4K6	4	28.10
Combo-5	Rpr4K ,Base4K1,Base4K5, Base4K6	4	28.17
Combo-6	Finetune1, Finetune2, Finetune3,Finetune4	4	27.31
Combo-7	Bagging1, Bagging2, Bagging3, Bagging4	4	27.40
Combo-8	Combo-1 , Combo-3	8	28.22
Combo-9	Combo-4,Dropout4,Base4K2,Base4K3,Base4K4	8	28.32
Combo-10	Combo-5,Base5,Big,Base4K4,Dropout3	8	28.56
Combo-11	Combo-10,Dropout4,Base,Base4K2,Base4K3	12	28.59
Combo-12	Combo-11, ,Finetune4,Bagging1,Base4K7,Base4K8	16	28.48

3.5 多样性与复杂度分析

本节我们将介绍模型之间多样性的评价标准与 Ensemble 系统的时间复杂度开销。

多样性 模型间的多样性决定了多个模型进行融合后翻译性能的涨幅。本文通过计算不同模型预测出的译文与参考译文之间的 N-gram 分数来评估多样性的大小。我们使用 multi-bleu.perl 脚本，将多个模型的译文作为参考答案，并使用参考译文作为假设，计算模型间的 N-gram 相似度。如表 3 所示：我们混合不同单模型在测试集上的预测集合，发现当模型数量一定时，模型之间结构差异性越大，模型之间的多样性越高；此外，伴随着模型数量的增加，模型间的多样性增大，但增大的幅度越来越小。通过对比模型数量为 4 时的不同模型结构的前 3 组实验，我们发现使用 dropout 的多样性指标略低于其他两组，因此使用该设置进行预测结果融合后，模型性能的涨幅也低于其他两组，与 3.4.1 实验现象相吻合。

表 3 多样性分析

系统名称	Diversity
4 base	36.55
4 filter4096	36.97
4 dropout	35.88
4 mix	37.12
8 mix	40.97
12 mix	43.15
16 mix	44.02

复杂度 Ensemble 模型进行预测的时间复杂度在不考虑卡间参数传递损失的前提下，与单模型相近。但在真正的推断过程中，我们在预测每个词时都采用同步的方式来整合不同卡上模型计算出的概率分布向量，因此卡间的信息传递占用了很大的时间开销。此外，由于不同模型结构的计算量不同，会导致解码速度不统一，容易造成木桶效应。根据对不同数量的模型进行 Ensemble 实验，伴随着数量的增加，解码所需的时间也变长。因此提高卡间的传输速率会大幅度提升集成学习在实际应用中的效率。

3.6 整合实验结果

在表 4 中我们展示了模型参数平均与预测结果融合的最好结果，实验结果表明对比 Transformer_base 单模型,参数平均的方法提升了近 0.8 个 BLEU 值，在参数平均的基础上融合 12 个模型得到 3.19 个 BLEU 值提升。

表 4 WMT17 中英测试集结果

系统名称	BLEU
基线	25.4
参数平均	26.22(+0.82)
Ensemble12	28.59(+3.19)

4 相关工作

最近神经机器翻译越来越受外界的关注，现有的大多数工作都是针对模型结构的改进与经验性方法的集成。近年来出现很多在解码端融入不同手段来启发式的改进翻译性能的工作，常见的手段有将集成学习的思想融入翻译推断过程[8-11]；融合最小贝叶斯风险到 NMT[24]；使用不同系统得到有差异性的翻译结果，通过系统融合的方式利用混淆网络来重构翻译结果[25]等等。

本文重点关注集成学习在神经机器翻译中的应用，致力于探索一种更有效的融合方式。近期 Vaswani 等人提出了基于自注意力机制的 Transformer 模型颠覆了研究人员的认知，更快的模型训练收敛时间与更显著地性能优势让 Transformer 代替了循环神经网络成为了最受欢

迎的端对端翻译模型。在文章中，Vaswani 提出了对最后 N 个模型保存点进行参数平均可以得到方差更小、性能更强的单模型，但并没有详细地介绍如何设置模型的保存间隔与参数平均模型的个数。Sennrich 等人在 WMT16 比赛[8]中第一次提出对单一模型的最后 4 个保存的模型进行参数平均，并且在 WMT17 比赛[9]中尝试用不同的超参数训练了 4 个不同结构模型，在中英任务上照比基线有+1.5BLEU 的提升。但他们并没有详细介绍修改超参数的策略从而获得差异性更大的模型。另一方面他们只采用了 4 个模型进行融合，没有尝试更多的模型是否能进一步提升翻译的性能。搜狗[10]与厦门大学[11]均在 WMT17 的中英与英中比赛使用了预测结果融合的手段取得了性能的提升，但同样也没有给出融合更多样模型的实验结论。除此之外，Freitag M[26]等人使用老师-学生（Teacher-Student）框架让网络结构更简单的单模型去逼近 Ensemble 模型的性能，同样并没详细介选取模型的策略。

本文通过大量的实验对比分析集成学习在 NMT 中的应用方法，针对模型与数据多样性，特别在更多模型融合方面给出了经验性结论，翻译性能得到显著地提升。

5 结论

本文的主要贡献在于结合模型参数平均与预测结果融合两种集成学习在 NMT 中的应用手段，在大规模语料上进行实验，总结了一种更高效的集成方法。一方面我们发现经过参数平均后的模型有更强的表示能力；另一方面实验结果表明在融合模型数量相同的情况下，更加多样性的模型组合会在性能上带来更大的提升，尤其在融合更多模型的角度进一步实验，发现仍然有较大的提升空间。本文在 WMT17 中英测试集上，选用 12 个经过参数平均 15 后的多样性子模型，对比于基线提高了近+3.2BLEU。我们的方式证实了更多更具差异性的模型进行融合能显著地提升翻译性能，同时对模型多样性与解码复杂度进行了系统的分析。

6 未来工作

未来我们会在更多的语种方向来验证 Ensemble 实验结论的有效性；另一方面模型融合系统对于超参数长度惩罚 alpha 非常敏感，我们将从句子级 BLEU 的角度分析不同 alpha 对翻译结果的影响，并在解码每一句之前用额外的系统去预测 alpha 值，进而提升翻译的性能。

参考文献

[1] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001.
[2] Dietterich T G. Ensemble Methods in Machine Learning[J]. multiple classifier systems, 2000: 1-15.
[3] Bauer E J, Kohavi R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants[J]. Machine Learning, 1999: 105-139.
[4] Opitz D W, Maclin R. Popular ensemble methods: an empirical study[J]. Journal of Artificial Intelligence Research, 1999, 11(1): 169-198.
[5] Zhou Z, Jiang Y. Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble[J]. international conference of the ieee engineering in medicine and biology society, 2003, 7(1): 37-42.
[6] Xu L, Krzyzak A, Suen C Y, et al. Methods of combining multiple classifiers and their applications to handwriting recognition[J]. systems man and cybernetics, 1992, 22(3): 418-435.
[7] Xiao T, Zhu J, Liu T, et al. Bagging and Boosting statistical machine translation systems[J]. Artificial Intelligence, 2013: 496-527.
[8] Sennrich R, Haddow B, Birch A. Edinburgh Neural Machine Translation Systems for WMT 16[J]. 2016:371-376.
[9] Sennrich R, Birch A, Currey A, et al. The University of Edinburgh's Neural MT Systems for WMT17[J]. 2017.
[10] Wang Y, Cheng S, Jiang L, et al. Sogou Neural Machine Translation Systems for WMT17[C]// Conference on Machine Translation. 2017:410-415.
[11] Tan Z, Wang B, Hu J, et al. XMU Neural Machine Translation Systems for WMT 17[C]// Conference on Machine Translation. 2017:400-404.
[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[J]. neural information processing systems, 2017: 5998-6008.
[13] Zhou Z, Shin J Y, Zhang L, et al. Fine-Tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally[C]. computer vision and pattern recognition, 2017: 4761-4772.
[14] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40(2):139–157, 2000b.
[15] Bahdanau D, Cho K, Bengio Y, et al. Neural Machine Translation by Jointly Learning to Align and Translate[J]. international conference on learning representations, 2015.
[16] Sutskever I, Vinyals O, Le Q V, et al. Sequence to Sequence Learning with Neural Networks[J]. neural information processing systems, 2014: 3104-3112.
[17] Ba J L, Kiros J R, Hinton G E. Layer Normalization[J]. 2016.
[18] Tong Xiao, Jingbo Zhu, Hao Zhang, and Qiang Li. 2012. Niutrans: an open source toolkit for phrase-based and syntax-based machine translation. In Proceedings of the

ACL 2012 System Demonstrations, pages 19–24. Associationfor Computational Linguistics.
[19] Moore R C, Lewis W D. Intelligent Selection of Language Model Training Data[C]. meeting of the association for computational linguistics, 2010: 220-224.
[20] Sennrich R, Haddow B, Birch A, et al. Improving Neural Machine Translation Models with Monolingual Data[J]. meeting of the association for computational linguistics, 2016: 86-96.
[21] Koehn P, Hoang H, Birch A, et al. Moses: Open Source Toolkit for Statistical Machine Translation[C]. meeting of the association for computational linguistics, 2007: 177-180.
[22] Sennrich R, Haddow B, Birch A, et al. Neural Machine Translation of Rare Words with Subword Units[J]. meeting of the association for computational linguistics, 2016: 1715-1725.
[23] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[J]. international conference on learning representations, 2015.
[24] Shu, R., Nakayama, H.: Later-stage minimum bayes-risk decoding for neural machine translation (2017)
[25] Zhou, L., Hu, W., Zhang, J., Zong, C.: Neural system combination for machine translation. In: Meeting of the Association for Computational Linguistics. pp. 378–384 (2017)
[26] Freitag M, Alonaizan Y, Sankaran B. Ensemble Distillation for Neural Machine Translation[J]. 2017.

李北（1995—），硕士研究生，主要研究领域为机器翻译。
E-mail：libei_neu@outlook.com

王强（1990—），博士研究生，主要研究领域为机器翻译。
E-mail：wangqiangneu@gmail.com

姜雨帆（1994—），硕士研究生，主要研究领域为机器翻译。
E-mail：jiangyufan2018@outlook.com